Flexible modelling of vaccine effect in self-controlled case series models

Flexible modelling of vaccine effect in self-controlled case series models

Yonas Ghebremichael-Weldeselassie, Heather J. Whitaker and C. Paddy Farrington

Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.

†

Summary.
The self-controlled case-series method (SCCS), commonly used to investigate the safety of vaccines, requires information on cases only and automatically controls all age-independent multiplicative confounders, while allowing for an age dependent baseline incidence.
Currently the SCCS method represents the time-varying exposures using step functions with pre-determined cut-points. A less prescriptive approach may be beneficial when the shape of the relative risk function associated with exposure is not known a priori, especially when exposure effects can be long-lasting. We therefore propose to model exposure effects using flexible smooth functions. Specifically, we used a linear combination of cubic M-splines which, in addition to giving plausible shapes, avoids the integral in the log-likelihood function of the SCCS model. The methods, though developed specifically for vaccines, are applicable more widely.
Simulations showed that the new approach generally performs better than the step function method. We applied the new method to two data sets, on febrile convulsion and exposure to MMR vaccine, and on fractures and thiazolidinedione use.

Keywords: M-splines; Risk function; Self controlled case series; Smoothing; Vaccines.

1. Introduction

The self controlled case series (SCCS) method is an epidemiological study design used to assess the association between time varying exposures and an adverse event of interest (Farrington, 1995). In the standard SCCS framework, exposure histories are collected for cases, namely individuals who experienced the event of interest at least once, over a defined period during which individuals are observed (the observation period). Appropriate conditioning enables an unbiased estimate of the relative incidence of the event to be obtained, this relative incidence being the ratio of the incidence rate in a predefined post-exposure risk period to the incidence rate at other times (the control period) within the observation period. The method implicitly controls for all measured and non-measured time independent confounding variables that act multiplicatively on the baseline incidence rate, but time-varying confounding variables should be modelled explicitly.

There has been much work on flexible ways of modelling the exposure effect for standard study designs. These involve representing the exposure history as a convolution of past

†*Address for correspondence:* Yonas Ghebremichael-Weldeselassie, Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.
E-mail: yonas.weldeselassie@open.ac.uk

exposures that combines information about duration, intensity and timing of exposure in one summary measure, as proposed by Breslow et al. (1983) and Thomas (1988). Letting $z(u)$ to be dose or intensity of exposure at age u and $w(u,t)$ a function that assigns weights to past exposures, the weighted cumulative exposure (WCE) at time t is defined as

$$WCE(t) = \int_0^t z(u)w(u,t)du. \tag{1}$$

Within this context, interest has focused on modelling the weight function $w(u,t)$, whether by a priori parametric models (Vacek, 1997; Langholz et al., 1999; Abrahamowicz et al., 2006) or spline models of varying complexity (Hauptmann et al., 2000, 2001; Berhane et al., 2008; Sylvestre and Abrahamowicz, 2009), with applications to environmental and drug exposures.

The focus of the present paper is the representation of the relative incidence associated with exposure within SCCS vaccine studies. In the case of vaccines, a point exposure (corresponding to the administration of antigen) occurs at the age of vaccination c, so $z(u)$ is a Dirac delta function. Setting $w(u,t) = w(t-u)$ we obtain the WCE function

$$WCE(t) = w(t-c) \text{ for } t > c, \text{ 0 otherwise.} \tag{2}$$

While our focus is on vaccines, the approach we develop has broader applicability, as will be shown in one of our examples. In current SCCS methodology, $WCE(t)$ is represented by a step function, with pre-determined cut-points. This is not biologically plausible and may incur losses in efficiency (Greenland, 1995; Weinberg, 1995; Zhao and Kolonel, 1992). Furthermore, epidemiologists frequently find it difficult to select specific cut points a priori and a poor choice of cut-points may be associated with cut-point bias and misclassification (Altman, 1991; Greenland, 1995). Therefore, we represent the exposure-related relative incidence function (which is a function of time since exposure) as a linear combination of cubic M-spline basis functions, which are variants of B-splines, thus replacing the step function with a smooth flexible function.

The paper is organized in six sections. Section 2 briefly introduces the likelihood function of the SCCS model. Section 3 describes an analysis of data on febrile convulsion and MMR vaccine using the standard SCCS method. This is followed by the representation of the exposure-related relative incidence function as a linear combination of cubic M-splines in Section 4. Section 4 also includes analysis of the data on febrile convulsion and MMR vaccine using the new method. Application of the new method to a non-vaccine study that investigates the potential association between fracture and thiazolidinedione use is also presented. Section 5 presents a simulation study conducted to evaluate the performance of the new method and to compare it with the existing step function approach. In Section 6 we make some final remarks.

2. The case series likelihood

Suppose that cases, indexed by i, $i = 1, ..., N$, are observed from age a_i to age b_i and experience a vector of exposures $x_i(t)$ at age t within the observation period. Being a case means that at least one event occured in $(a_i, b_i]$. Events are assumed to arise with rate $\lambda_i(t|x_i^t)$ where $x_i^t = \{x_i(s) : s \leq t\}$ represents the exposure history of individual i up to age t. Thus $x_i^{b_i}$ is the entire exposure history of individual i up to the end of their observation period.

The SCCS conditional likelihood is obtained by conditioning on the number of events, n_i, experienced by an individual i during their observation period $(a_i, b_i]$. Three assumptions are required: (1) events arise in a non-homogenous Poisson process; (2) $\lambda_i(t|x_i^t) = \lambda_i(t|x_i^{b_i})$, which implies that the exposure histories to the end of the observation period must be independent of the occurrence of an event at time t; and (3) censoring of individuals at the end of the observation period occurs completely at random, so that the occurrence of the event of interest must not censor or affect the observation period (Farrington, 1995; Farrington and Whitaker, 2006; Whitaker et al., 2006; Weldeselassie et al., 2011). Departures from these assumptions are discussed in Farrington et al. (2009, 2011).

Suppose that the event intensity is parameterized as a proportional incidence model of the form

$$\lambda_i(t|x_i^t) = \varphi\psi(t)\exp\left\{\gamma_i + x_i(t)^T\beta\right\},$$

where φ is the underlying incidence at some reference age, γ_i is a sum of fixed and random individual effects, and $\psi(t)$ is the age-specific relative incidence function. The focus of inference is the parameter β. The SCCS conditional likelihood function is then given by

$$L = \prod_{i=1}^{N}\prod_{j=1}^{n_i}\frac{\psi(t_{ij})\exp\left\{x_i(t_{ij})^T\beta\right\}}{\int_{a_i}^{b_i}\psi(t)\exp\left\{x_i(t)^T\beta\right\}dt}, \tag{3}$$

where t_{ij} is age at the j^{th} event for individual i. From this we can see that SCCS has two major features: (1) it automatically controls for all time-independent confounding covariates that act multiplicatively (since these cancel out), and (2) only cases (individuals with at least one event) need to be included in the analysis (since terms with $n_i = 0$ contribute 1 to the likelihood).

In the standard SCCS method the age-specific relative incidence function $\psi(t)$ and the exposure-related relative incidence function $\exp\left\{x_i(t)^T\beta\right\}$ are represented by piecewise constant step functions.

3. Standard SCCS analysis

In this section we analyse a data set on febrile convulsion and MMR vaccine using the standard SCCS method where age and exposure effects are represented using step functions. Different exposure groups will be used to investigate the association between the outcome event and exposure.

The aim of the analysis is to investigate the association between febrile convulsions and MMR vaccine. Febrile convulsions or seizures are a relatively common childhood conditions. MMR is a combined vaccine that protects against measles, mumps and rubella (German measles).

The data on febrile convulsions and paediatric vaccines were collected in England and Wales in the period 1991-1994. The data set includes 2,389 children aged 28-730 days, who had 3,826 febrile convulsion events. Of the 2,389 children, 2,021 were vaccinated with MMR. The average age at which MMR vaccine was administered was 437 days. Children with age at event outside the age interval $28 - 730$ days were excluded from the analysis.

The exposure risk periods and age groups were represented by piecewise constant functions. For the exposure effect we chose three different categorizations: (1) 10 exposure risk periods between 0 and 50 days with cut points at $6, 11, 18, 22, 26, 30, 36, 40$ and 45 days since vaccination, (2) three risk periods $(0, 11]$, $(11, 30]$ and $(30, 50]$ days and (3) three risk

periods $(0, 25]$, $(25, 50]$ and $(50, 75]$ days since vaccination. For all the three analyses we used 21 age groups of length 30 days while the first and last groups were of length 32 and 40 days respectively to include the age effect in the model. Results of these analyses are presented in Table 1

[Table 1 about here.]

From analysis (1) in Table 1, it can be seen that there is a significant association between febrile convulsion and MMR vaccine in the risk periods $(6, 11]$ and $(18, 22]$ days since vaccination. In the second analysis where there are only three risk periods the increased risk of febrile convulsion is only detected in the first risk period of $0 - 11$ days and the relative incidence estimate obtained in this analysis, 2.25(1.91, 2.65), is much less than the value obtained in the first analysis, 3.49(2.88, 4.23). From the third analysis, there is an increased risk of febrile convulsion in the risk periods $(0, 25]$ and a borderline significant relative incidence in the risk period $(25, 50]$. The relative incidence in the risk period $(50, 75]$ was found to be non significant.

These results show that different categorizations of the risk periods lead to different results. Therefore, in the absence of prior hypotheses about the risk period, a new way of modelling exposure effect that does not have this limitation is required. Conversely, even if prior hypotheses about the risk period are available, it is of interest to estimate the post-vaccination risk function over a wider time period. We propose to use spline functions, namely linear combination of cubic M-splines.

4. Smooth exposure effect

In this paper we approximate the exposure-related relative incidence function by spline functions. This allows us to provide smooth estimates with continuous first two derivatives. Splines are flexible enough to represent a variety of clinically plausible shapes (Smith, 1979). To begin with, we specify a nominal maximum risk period over which the exposure-related relative incidence function can be different from 1; outside this interval (which may be unbounded to the right), the function will take the value 1. The argument of this function is time since start of exposure (in our case, vaccination).

The exposure-related relative incidence function is required to be a positive function. Therefore, we use a linear combination of M-spline basis functions, which are variants of B-splines. An M-spline of order q is thus a positive function constructed by combining pieces of polynomial functions of degree $q - 1$ connected at knots (Ramsay, 1988; Ghebremichael-Weldeselassie *et al.*, 2014). To keep positivity of the M-splines when combined linearly, we constrain their coefficients to be positive. Therefore, the function representing the exposure effect in equation (3), $\exp\left\{x_i(t_{ij})^T\beta\right\}$, will be replaced by a function of time since exposure represented as a linear combination of M-splines of order 4 (i.e. cubic splines):

$$\omega(t-c) = \begin{cases} \sum_{l=1}^{m} g(\beta_l) M_l(t-c), & c < t \leq d \\ 1, & \text{otherwise,} \end{cases}$$

where $g(\beta_l)$ are parameters to be estimated that determine the shape of the function, c is age at start of exposure, d is age at end of the nominal risk period and m is the number of M-spline functions. We shall choose $g(\beta_l) = \beta_l^2$ to ensure positivity of the function. $g(\beta_l) = \exp(\beta_l)$ can also be used but may have a convergence problem when $g(\beta_l)$ should be zero. The value m depends on the number of interior knots and the order of M-splines

chosen: m = number of interior knots + order. Usually a number of interior knots between 8 and 12 is sufficient (Joly et al., 1998). We choose equidistant knots between 0 and the maximum of $d_i - c_i$ (where c_i and d_i are the beginning and end of the exposure-related risk period for individual i, so their difference represents the length of the nominal risk period for point exposures like vaccines), inclusive, and add an extra $q-1$ equidistant knots below the minimum and above the maximum knots to construct the M-spline basis functions. When risk periods are of indefinite length ($d = \infty$) d_i is set equal to the value of b_i.

Replacing the exposure effect in equation (3) by a linear combination of cubic M-splines, the log-likelihood function is

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{\psi(t_{ij})(\sum_{l=1}^{m}\beta_l^2 M_l(t_{ij}-c_i))^{I(c_i<t_{ij}\le d_i)}}{\int_{a_i}^{b_i}\psi(t)(\sum_{l=1}^{m}\beta_l^2 M_l(t-c_i))^{I(c_i<t\le d_i)}dt}\right). \tag{4}$$

The age-specific relative incidence $\psi(t)$ is represented by a step function, as in the standard SCCS method; age effects are usually not of primary interest and are generally more gradual. Thus, we subdivide the observation period of each case into intervals $(l_{ih}, u_{ih}]$, h indexing the age group, with age-specific relative incidence $\exp(\alpha_h)$. Without loss of generality, we can choose these intervals to be sufficiently narrow (by splitting them) that they are properly contained in $(c_i, d_i]$ or its complement in $(a_i, b_i]$. The log-likelihood is then:

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{\exp(\alpha_{h(i,j)})(\sum_{l=1}^{m}\beta_l^2 M_l(t_{ij}-c_i))^{I(c_i<t_{ij}\le d_i)}}{\sum_h \exp(\alpha_h)\int_{l_{ih}}^{u_{ih}}(\sum_{l=1}^{m}\beta_l^2 M_l(t-c_i))^{I(c_i\le l_{ih}<d_i)}dt}\right), \tag{5}$$

where $h(i,j)$ is the age interval containing t_{ij}.

The integral in the denominator of the log-likelihood function (5) can be replaced by a linear combination of integrated splines (I-splines) since the integral of an M-spline function of order q can be expressed as an I-spline of order $q+1$ (Ramsay, 1988). Hence, denoting the length of interval h for the i^{th} individual by $e_{ih} = u_{ih} - l_{ih}$, our log-likelihood function will be:

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{\exp(\alpha_{h(i,j)})(\sum_{l=1}^{m}\beta_l^2 M_l(t_{ij}-c_i))^{I(c_i<t_{ij}\le d_i)}}{\sum \exp(\alpha_h)(e_{ih})^{(1-I(c_i\le l_{ih}<d_i))}(\sum_{l=1}^{m}\beta_l^2 I_l(u_{ih}-c_i) - \sum_{l=1}^{m}\beta_l^2 I_l(l_{ih}-c_i))^{I(c_i\le l_{ih}<d_i)}}\right). \tag{6}$$

To estimate the parameters of interest from the log-likelihood (6), we introduce a penalty term that controls the smoothness of the exposure-related relative incidence function. As in O'Sullivan (1988) the penalty is based on the second derivative of the linear combination of cubic M-splines. Thus, the penalized log-likelihood function is:

$$\begin{aligned} pl &= l - \lambda\int(\sum_{l=1}^{m}\beta_l^2 M_l''(u))^2 du \\ &= l - \lambda((\boldsymbol{\beta}^2)^T \mathbf{A}\boldsymbol{\beta}^2) \end{aligned} \tag{7}$$

where l is the log-likelihood in (6), $\boldsymbol{\beta}^2$ is the vector with elements β_l^2, $\mathbf{A}$ is an $m \times m$ matrix with (r,l) element $\int M_r''(u)M_l''(u)du$ and $\lambda \ge 0$ is a smoothing parameter that controls the balance between smoothness of the function and fit to the data. One can also use a difference penalty as in Eilers and Marx (1996).

4.1. Smoothing parameter selection

We choose the smoothing parameter by maximizing an approximate cross-validation score, as proposed by O'Sullivan (1988), while keeping the age effect to be constant (that is setting the $\alpha_h = 0$).

As before, let $\boldsymbol{\beta}$ be the vector of parameters β_l. Denote the cross-validation score $V(\lambda)$,

$$V(\lambda) = \sum_{i}^{N} l_i(\hat{\boldsymbol{\beta}}_{-i})$$

where $\hat{\boldsymbol{\beta}}_{-i} = \hat{\boldsymbol{\beta}}_{-i}(\lambda)$ is the maximum penalized likelihood estimator of $\boldsymbol{\beta}$ (with $\alpha = 0$) when individual i is removed, and l_i is the log likelihood contribution of individual i. Following O'Sullivan (1988), $V(\lambda)$ may be approximated by $\bar{V}(\lambda)$,

$$\bar{V}(\lambda) = l(\hat{\boldsymbol{\beta}}) - \mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{H}), \tag{8}$$

where $\mathrm{tr}(X)$ is trace of a matrix X, $\hat{H}$ is the likelihood component of the Hessian evaluated at the penalized MLE, $\hat{\boldsymbol{\beta}}$, and $2\lambda\mathbf{S}$ is the penalized component of the Hessian. The matrix $\mathbf{S}$ depends on $g(\beta_l)$. In our case $g(\beta_l) = \beta_l^2$, therefore $\mathbf{S} = 4\left(\mathbf{A} \circ (\boldsymbol{\beta}\boldsymbol{\beta}^T)\right) + 2(\mathrm{diag}(\mathbf{A}\boldsymbol{\beta}^2))$ (see Ghebremichael-Weldeselassie *et al.* (2014)). The symbol $\circ$ denotes pointwise matrix multiplication. If $g(\beta_l) = \beta_l$ then $\mathbf{S} = \mathbf{A}$ (Joly et al., 1998).

The penalized log likelihood function (7) with no age effect is maximized for a grid of λ values, and the value of λ that maximizes the approximate cross-validation score is used in a final optimization step with the full model to obtain the relative incidences related to age and the relative incidence function related to exposure.

4.2. Approximate variability bands

Following O'Sullivan (1988) and Joly et al. (2002), we use a Bayesian-like technique to generate variability bands for the exposure-related relative incidence estimators. Considering the penalized log-likelihood function (7) to be a posterior log-likelihood for $\boldsymbol{\beta}$ and the penalty term to be a prior log-likelihood, the approximate covariance of $\hat{\boldsymbol{\beta}}$ is $\hat{V}_{pl}$, where $\hat{V}_{pl}$ is the negative of the inverted hessian of pl evaluated at the penalized maximum log-likelihood estimates. Our approximation of the exposure-related relative incidence function used $g(\beta_l) = \beta_l^2$ to keep positivity of the function, we therefore need to know the covariance of $\boldsymbol{\beta}^2$. The required covariance matrix can be obtained using the delta method as $\hat{V}_{tr} = 4\mathrm{diag}(\hat{\boldsymbol{\beta}})[\hat{V}_{pl}](\mathrm{diag}(\hat{\boldsymbol{\beta}}))^T$. Hence the approximate 95% variability bands for the exposure-related relative incidence are

$$\hat{\omega}(\tau) \pm 1.96\sqrt{M(\tau)^T\hat{V}_{tr}M(\tau)}$$

where τ is time since start of exposure and $M(\tau)^T = (M_1(\tau), \ldots, M_m(\tau))$.

Alternatively, to ensure that the variability bands lie above zero, they can be obtained on the log scale as

$$\hat{\omega}(\tau)\exp\{\pm 1.96\sqrt{M(\tau)^T\hat{V}_{tr}M(\tau)}/\hat{\omega}(\tau)\}.$$

4.3. Analysis of febrile convulsion and MMR vaccine

We now apply the new method to investigate the association between febrile convulsion and MMR vaccine. In this analysis we used 50 days post MMR vaccine to represent the exposure effect with splines. Since for point exposures all individuals have the same nominal risk period of 50 days, we defined 12 equidistant inner knots between 0 and 50 days. Age was included in the model as a step function with the same age groups used in the standard SCCS analysis in Section 3. A linear combination of cubic M-splines was used to represent the MMR-related relative incidence function. The value of the smoothing parameter selected by the approximate cross-validation score was 0.031. We present the relative incidence function estimated by maximizing the penalized log-likelihood function (7) along with its approximated variability bands in Panel (a) of Figure 1. The figure shows no risk of febrile convulsion in the first 3 days post MMR vaccination and a borderline non-significant relative incidence of 1.248 at the 4^{th} day. However, there is a significantly increased risk between 5 and 11 days after exposure to the vaccine. The relative incidence at the 5th day is 1.922 and increases smoothly to 3.647 at the 8th day and then the risk decreases to 1.244 at 12 days since exposure. There is also an increased risk of febrile convulsion due to MMR vaccine between 19 and 21 days post vaccination. At all other times after vaccination there is no significantly increased risk of febrile convulsion. We also did the same analysis with a nominal risk period post MMR vaccination of 75 days as presented in Panel (b) of Figure 1. This resulted in a similar relative incidence function.

[Figure 1 about here.]

Figure 2 compares the effects of exposure to MMR vaccine estimated by the standard SCCS in Section 3 and the spline based methods. The standard model presented here is that used in analysis (1) of Section 3 with exposure cut points at $6, 11, 18, 22, 26, 30, 36, 40$ and 45 days since vaccination and 21 age groups. The results given by the two methods are similar, however the result from the spline method is different from the other two analyses in Section 3.

[Figure 2 about here.]

4.4. Analysis of fractures and thiazolidinediones

The methods developed in the present paper can be applied more widely. We illustrate this with data on fractures and thiazolidinediones, which were previously analysed by Douglas et al. (2009) using the standard case series method. The aim of the study was to investigate whether there is an increased risk of fracture associated with the use of thiazolidinediones, a class of medicines used to treat type 2 diabetes. The data used in the analysis were primary care computerized clinical records from the United Kingdom-based General Practice Research Database (GPRD). 1819 patients aged about 40 years or older prescribed at least one thiazolidinedione and with at least one fracture were included in the analysis. The data included patients with multiple fractures: 283 (16%), 64 (4%), and 25 (1%) had two, three, and four or more fractures, respectively. Multiple fractures were included in the analysis if the fractures happened at different sites or at the same site but at least 6 months apart.

In Douglas et al. (2009) the authors defined the control period to be from age at start of observation period until age at first prescription of a thiazolidinedione and the risk period was from age at start of thiazolidinedione use until age at end of observation period. The length of exposure following each individual prescription was calculated using information

recorded in the GPRD on pack size and dosing frequency. Thiazolidinedione treatment was assumed to be continuous where any apparent treatment break was less than 60 days, to allow for partial noncompliance and situations where patients may have built up treatment stocks (Douglas et al., 2009). Age at end of observation was then taken to be age at the earliest of any treatment break longer than 60 days or the end of recorded follow up in the database. The mean duration of control periods prior to thiazolidinedione use was 9.5 years, and the mean duration of exposure to a thiazolidinedione was 2.3 years.

Unlike vaccines, thiazolidinediones are not point exposures, however we can use a similar approach as with vaccines by taking $z(u) = z$ for $u > c$, the age at first thiazeledinedione, and 0 otherwise, and $w(u,t) = w(t-u)$ in equation (1). Letting $W(x)$ denote the integral of $zw(x)$, we then have $WCE(t) = W(t-c)$ as in equation (2). This leads to the same likelihood function as before (6). In this model, the intensity of exposure is assumed constant after initiation at time c, and $WCE(t)$ just reflects the effect of exposure duration.

We reanalyzed the data using the new version of SCCS where time since exposure is represented by a linear combination of M-splines. The maximum duration of exposure to thiazolidinedione was 2364 days. Hence our exposure-related relative incidence function was represented by a linear combination of cubic M-splines defined between 0 and 2364 days since first exposure. We chose 14 equidistant knots between 0 and 2364 days inclusive, that is we have 16 M-spline basis functions. The time-varying confounding covariate age was taken into account using a piecewise constant function with 42 age groups: the first age group is less than 14610 days (40 years) of age, followed by 5 age groups of length two years, 28 groups of 1 year length, 7 groups of length two years and the last age group with age greater than 33603 days (92 years).

To estimate the parameters we first selected the optimum smoothing parameter, λ, that maximizes the approximate cross-validation score in equation (8). This optimum λ was 288. We then maximized the penalized log-likelihood function in (7) for fixed $\lambda = 288$ to get the required parameter estimates. The estimated exposure relative incidence function and its approximate variability bands are presented in Figure 3.

[Figure 3 about here.]

From panel (a) of Figure 3, it can be seen that the relative incidence of fracture due to thiazolidinedione use increases initially as time since exposure increases. There is no significant increased risk of fracture in the first two months of exposure and the relative incidence is borderline significant from two months to about 1 year and half, but there is a significantly increased risk of fracture due to exposure to thiazolidinedione thereafter, and the maximum relative incidence of 2.103 is reached after about 5 years of exposure. The relative incidence may start to decrease and the variability bands widen after 5 years.

In their parametric SCCS analysis, Douglas et al. (2009), defined five exposure groups of $(0,1], (1,2], (2,3], (3,4]$ and $(4,7]$ years since first exposure and obtained relative incidence estimates of $1.26, 1.49, 1.70, 2.31$, and 2.00 respectively. We repeated the analysis but with a different number and length of exposure groups, motivated by the spline model. We divided the time since first exposure in to 13 groups of lengths 6 to 9 months. Results from this analysis are presented in panel (b) of Figure 3 and are similar to those obtained by the spline method. The results obtained in Douglas et al. (2009) are slightly different from the results obtained using the new method.

[Figure 3 about here.]

5. Simulation study

To evaluate the performance of the new approach and compare it with the standard SCCS model, we conducted a simulation study. The number of cases was fixed at 1000. The length of the observation period for all cases was chosen to be 730 days, where age at start of observation $a_i = 0$ days and age at end of observation $b_i = 730$ days for all cases. Ages at vaccination c_i were generated within $(0, 730]$ from a uniform distribution and an exponential density with rate 0.003. Observation periods in vaccine safety studies are typically 1-2 years and the 2-year observation period selected here reflects this; for example, a study on influenza vaccine will have an observation period covering one or two flu seasons in calendar time, or a study on a routine childhood vaccine will perhaps include two years of age covering the age at which the vaccine is scheduled and some time after.

Six different scenarios of true exposure-related relative incidence functions were considered, four of them generated from beta densities and the other two from step functions with seven and three intervals respectively (Figure 4). The risk periods considered in all scenarios were of length 49 days. The effect of age was represented using a step function in which we used 6 equal age groups with true relative incidence rates $1, 2, 5, 8, 10$ and 15. We also considered a scenario where the age effect is represented by a continuous function with age-specific relative incidences generated from $8(\sin(0.01 \times t)) + 9$.

[Figure 4 about here.]

Marginal numbers of events per individual were generated from a Poisson distribution truncated to exclude 0 counts. A multinomial distribution was used to identify in which intervals within the observation period the events occurred and then a uniform distribution was used to generate event ages within these intervals. When the age effect is continuous, event ages are generated from the multinomial distribution with each day as distinct interval. For each scenario 100 samples of 1000 cases were generated in this way. These simulated data were then analysed using both the standard SCCS and the new approach with risk periods totalling 49 days following exposure (as simulated) or with an extended nominal risk period of 98 days. In the standard SCCS, the risk period of 49 days following an exposure was divided in to 7 groups of length 7 days (with 7 parameters). We also used an extended nominal risk period of 98 days, and fitted a standard SCCS model with 14 7-day groups (and 14 parameters). In addition, we fitted the standard SCCS model with 49-day risk intervals (and hence 1 or 2 parameters, according to the nominal risk period). For scenario 6 where the true exposure-related relative incidence is a step function with three intervals, the standard SCCS method was fitted with three exposure groups. In all the spline-based analyses we used 9 interior knots and the approximate cross-validation score was employed to choose the smoothing parameter.

[Figure 5 about here.]

Figure 5 shows the estimated exposure-related relative incidence curves obtained by fitting the spline-based method to the 100 randomly selected samples. The top row in the figure presents results obtained when the risk period is kept at 49 days post exposure (which is equal to the risk period used to simulate the data) and in the bottom row are the results when a nominal risk period of 98 days was used. The results show that the shapes of the true relative incidence curves (white lines) were captured well by most of the estimated curves.

[Table 2 about here.]

Table 2 shows that the mean integrated square errors (MISE) are all lower for the spline method than the standard method, except for scenario 2, in which the true exposure-related relative incidence was constant. For this scenario, the correctly specified step function model (with 1 or 2 parameters), though interestingly not the over-specified step function model (with 7 or 14 parameters), outperforms the spline model. Comparable if slightly degraded results were obtained for scenarios 1, 3 and 4 with the 98-day nominal risk period as with the correct 49-day risk period. For scenario 2, the spline method produced much worse results with 98 day compared to 49 day risk periods.

More simulation results are presented in Table 3, where the true age-specific relative incidence rate is a continuous function and the ages at exposure are generated from an exponential distribution. Similar to the results presented in Table 2 the spline method shows a better performance. The spline method gave a better performance even when the true exposure-related relative incidence function is represented by a step function of 7 groups as compared to the standard SCCS method with 7 exposure groups. However, when the true relative incidence is a step function with three groups the standard SCCS method is slightly better than the spline method. In both the spline and the standard SCCS methods the age effect was represented by a step function with 6 age groups. In addition to the mean and standard deviation of the integrated squared errors, 95% coverage probabilities of the true exposure-related relative incidences at 10, 25 and 45 days since the start of exposure are presented in Table 3.

Figure 6 shows the bias (top row) and variability (standard deviation, bottom row) of estimates from the standard (with 7 parameters) and spline-based SCCS methods with a 49 day nominal risk period. The bias of the standard method has a saw-tooth appearance in scenarios 1, 3 and 4 related to discontinuities at the cut-points, whereas the spline method occasionally shows some bias at endpoints, notably for scenarios 2 and 3. The spline method produces lower standard deviations, except at the endpoints.

[Figure 6 about here.]

6. Final remarks

In this paper we have proposed the use of regression splines to model the risk of an adverse event following vaccination, and showed how this might be applied to a drug-related exposure, in the self-controlled case series method. Specifically, we modelled the exposure-related relative incidence function using a linear combination of cubic M-splines.

We demonstrated how results differed when different exposure risk group cut-points were specified using the standard SCCS method for a study on the association between febrile convulsion and MMR vaccine. This provides some background to our motivation for the development of a new way of modelling the exposure effect that avoids the limitations of the step functions with pre-specified cut-points employed by the standard SCCS method.

Our spline-based SCCS method can be considered as a special case of weighted cumulative exposure models used in environmental epidemiology, which have also made good use of spline models (Hauptmann et al., 2000; Sylvestre and Abrahamowicz, 2009). These approaches have used information criteria to choose the knots in defining the B-spline basis functions. In our case, we intentionally selected a large number of knots and introduced a penalty term to the log-likelihood function to avoid over-fitting, the smoothing parameter

being chosen by an approximate cross validation score (O'Sullivan, 1988; Joly et al., 1998, 2002).

Simulation studies showed that the new approach generally has a better performance than the use of step functions in the context of the SCCS method. The new method was applied to two data sets to investigate the association between febrile convulsions and MMR, and between fracture and thiazolidinedione use. The estimates obtained from the new method are consistent with the results from the standard SCCS method when the exposure groups are well specified. Increasing the number of a priori defined exposure groups in the standard SCCS model may help in capturing the true exposure-related relative incidence curve better, but at the cost of reduced efficiency.

The new method will be especially useful in the absence of a clear, a priori hypothesis regarding the overall length of the exposure-related risk period. Consider the simple case of a standard SCCS study where there is a single risk period following an exposure: assuming that an association exists, there will be an optimal risk window starting at precisely the point in time when the exposure-related risk departs from the baseline risk and ending when this risk returns to the baseline. When using the standard SCCS method, if the a priori selected risk period is either too long or too short, estimates of relative incidence will be biased toward the null. This has been an issue of concern to epidemiologists using the standard SCCS method and Xu et al. (2011) proposed a method of identifying optimal risk windows for self-controlled case series studies of vaccine studies. However, this method assumed the relative risk to be constant throughout the risk periods, and hence could only identify a single risk window. In practice, several contiguous post-exposure risk periods are often used so periods of high and low risk can be identified. Our method offers several advantages. Only the start and end need be specified, and as the flexible function can allow for a exposure-related risk close to 0 (baseline), an overall risk window that is too long is allowed for. It is more efficient than using many contiguous risk periods as there are less parameters to estimate. It also enables a graphical risk profile that is not influenced by the choice of cut points to be obtained. This makes our method useful to explore biological hypotheses with an impact on the shape of the relative incidence curve. Or, if required, it can be used to help identify suitable cut points for multiple risk windows to be used in a standard SCCS analysis.

The new approach uses step functions to represent age effects as with the standard method. Usually age effects are not of primary interest and generally change more gradually. However, if the age effects are of interest one can represent the age-specific relative incidence function by a spline function as proposed by Ghebremichael-Weldeselassie *et al.* (2014) while keeping the exposure effect as a step function. Modelling both age and exposure effects using smooth functions at the same time would also be useful. To this end we are developing a non-parametric SCCS method where both effects are represented using spline functions.

While our focus has been on developing methods for studying the safety of vaccines, they have wider applicability, as we have shown in our example on fractures and thiazolidinediones. Further extension of the spline-based SCCS method to non-vaccine pharmacoepidemiology, notably to incorporate the effect of dose within a more general weighted cumulative exposure model framework, would be desirable.

Acknowledgements

We thank Ian Douglas (London School of Hygiene and Tropical Medicine) for the fracture data. We also thank the referees and editors for their useful comments. This research was supported by a Royal Society Wolfson research merit award to Paddy Farrington and MRC grant project HGXF A4468.

References

Abrahamowicz, M., Bartlett, G., Tamblyn, R., and du Berger, R. (2006) Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *Journal of Clinical Epidemiology,* **59(4),** 393–403.

Altman, D. G. (1991) Categorizing continuous variables. *British Journal of Cancer,* **64(5),** 975–975.

Berhane, K., Hauptmann, M., and Langholz, B. (2008) Using tensor product splines in modeling exposure-time-response relationships: application to the Colorado Plateau uranium miners cohort. *Statistics in Medicine,* **27,** 5484–5496.

Breslow, N.E., Lubin, J.H., Marek, P., and Langholz, B. (1983) Multiplicative models and cohort analysis. *Journal of the American Statistical Society,* **78(381),** 1–12.

de Boor, C. (1978) *A Practical Guide to Splines,*. New York: Springer-Verlag.

Douglas, I.J., Evans, S. J., Pocock, S., and Smeet, L. (2009) The Risk of Fractures Associated with Thiazolidinediones: A Self-controlled Case-Series Study. *PLoS Medicine,* **6(9),**.

Eilers, P. H. C., and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science,* **11,** 89–102.

Farrington, C. P. (1995) Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics,* **51,** 228–235.

Farrington, C. P., and Whitaker, H. J. (2006) Semiparametric analysis of case series data. *Journal of the Royal Statistical Society Series C-Applied Statistics,* **55,** 553–580.

Farrington, C.P., Whitaker, H. J., and Hocine, M. N. (2009) Case series analysis for censored, perturbed or curtailed post-event exposures. *Biostatistics,* **10,** 3–16.

Farrington, C.P., Anaya-Izquierdo, K., Whitaker, H. J., Hocine, M. N., Douglas, I., and Smeeth, L., (2011) Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association,* **106,** 417–426.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014). Self controlled case series method with smooth age effect. *Statistics in Medicine* **33(4),** 639 – 649.

Greenland, S. (1995) Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology,* **6(4),** 450–454.

Greenland, S. (1995) Dose-response and trend analysis in epidemiology - Alternatives to categorical analysis.*Epidemiology,* **6(4),** 356–365.

Hauptmann, M., Wellmann, J., Lubin J.H., Rosenberg, P.S., and Kreienbrock, L. (2000) Analysis of exposure-time-response relationships using a spline weight function. *Biometrics,* **56(4),** 1105–1108.

Hauptmann, M., Berhane, K., Langholz, B., and Lubin J.H. (2001) Using splines to analyse latency in the Colorado Plateau uranium miners cohort. *Journal of Epidemiology and Biostatistics,* **6,** 417–424.

Hauptmann, M., Pohlabeln, H., Lubin, J.H., Jckel, K.H., Ahrens, W., Brske-Hohlfeld, I., and Wichmann, H.E. (2002) The exposure-time-response-relationship between occupational asbestos exposure and lung cancer in two German case-control studies. *American Journal of Industrial Medicine,* **41,** 89–97.

Joly, P., Commenges, D., and Letenneur, L. (1998) A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics,* **54,** 185–194.

Joly, P., Commenges, D., Helmer C., and Letenneur L. (2002) A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics,* **3(3),** 433–443.

Langholz, B., Thomas, D., Xiang, A., and Stram, D. (1999) Latency Analysis in Epidemiologic Studies of Occupational Exposures: Application to the Colorado Plateau Uranium Miners Cohort. *American Journal of Industrial Medicine,* **35,** 246–256.

NHS (2013). http://www.nhs.uk/conditions/febrile-convulsions/pages/introduction.aspx *Accessed 23/06/2013.*

O'Sullivan, F. (1988) Fast computation of fully automated log-density and log-hazard estimators. *Siam Journal on Scientific and Statistical Computing,* **9,** 363–379.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Ramsay, J. O. (1988) Monotone Regression Splines in Action. *Statistical Science,* **3,** 425–461.

Smith, P.L. (1979) Splines as a useful and convenient statistical tool. *American Statistician,* **33(2),** 57–62.

Sylvestre, J., and Abrahamowicz, M. (2009) Flexible modeling of the cummulative effects of time-dependent exposures on the hazard. *Statistics in Medicine,* **28(27),** 3437–3453.

Thomas, D.C. (1988) Models for exposure-time-response relationships with applications to cancer epidemiology. *Annual Reviews of Public Health,* **9,** 451–482.

Vacek, P.M. (1997) Assessing the effect of intensity when exposure varies over time. *Statistics in Medicine,* **16,** 505–513.

Weinberg, C.R. (1995) How bad is categorization. *Epidemiology,* **6(4),** 345–347.

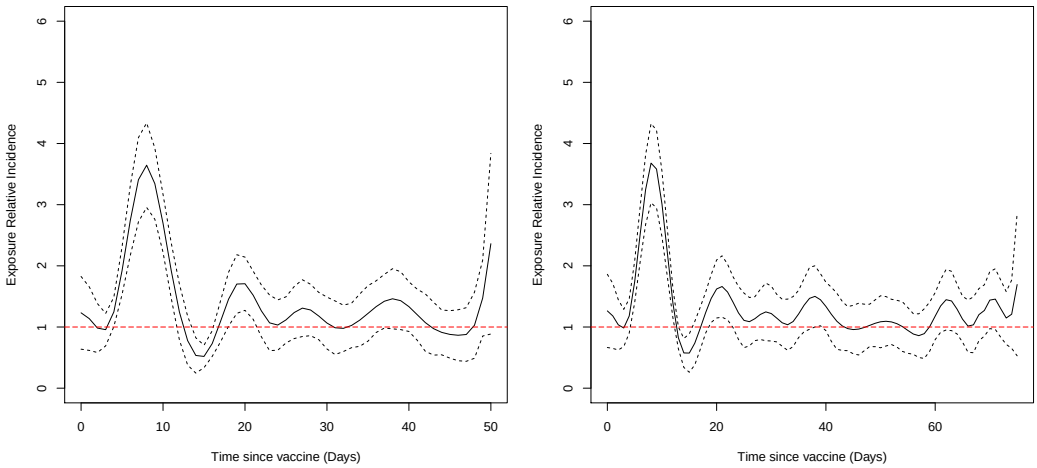
Weldeselassie, Y. G., Whitaker, H. J., and Farrington, C. P. (2011) Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. *Epidemiology and Infection,* **139,** 1805–1817.

Whitaker, H. J., Farrington, C. P., Spiessens, B., and Musonda, P. (2006) Tutorial in biostatistics: The self-controlled case series method. *Statistics in Medicine,* **25,** 1768–1797.

Whitaker, H. J., Hocine, M. N., and Farrington, C. P. (2009) The methodology of self-controlled case series studies. *Statistical Methods in Medical Research,* **18,** 7–26.

Xu, S.a , Zhang, L.a, Nelson, J.C.bc, Zeng, C.a, Mullooly, J.d, Mcclure, D.a, Glanz, J.a. (2011) Identifying optimal risk windows for self-controlled case series studies of vaccine safety. *Statistics in Medicine,* **30(7),** 742–752.

Zhao, L. P., and Kolonel, L. N. (1992) Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American Journal of Epidemiology,* **136(4),** 464–474.



(a) (b)

Fig. 1. Smooth estimates of the relative incidence function for exposure to MMR vaccine (solid lines) and 95% variability bands (dotted lines). Panel (a) shows the estimated function when the nominal risk period is 50 days post MMR vaccine and Panel (b) when it is 75 days.

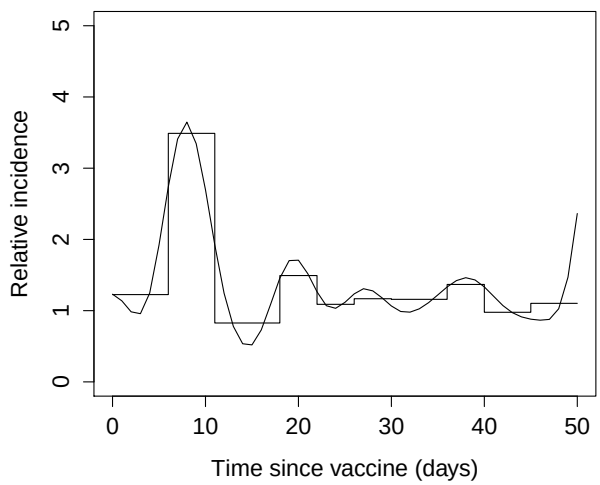
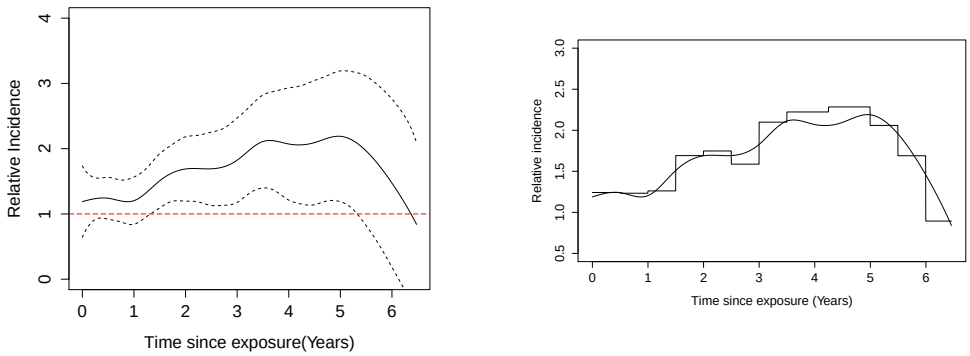


Fig. 2. Relative incidence functions for exposure to MMR vaccine estimated from the standard model with 10 exposure groups (step function) and spline-based SCCS (smooth function).



(a) (b)

Fig. 3. Panel (a) relative incidence function for thiazolidinedione use (solid line) and 95% variability bands (dotted lines). Panel (b) relative incidence functions estimated from spline-based method (smooth function) and standard SCCS (step function).

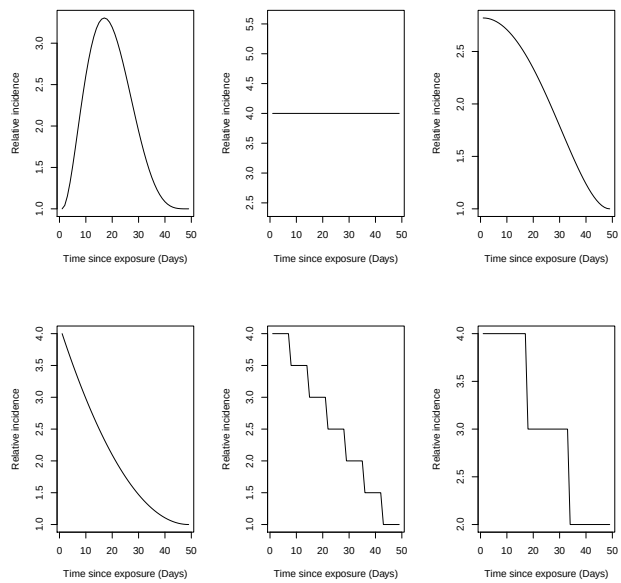


Fig. 4. True exposure-related relative incidence curves used in simulations.

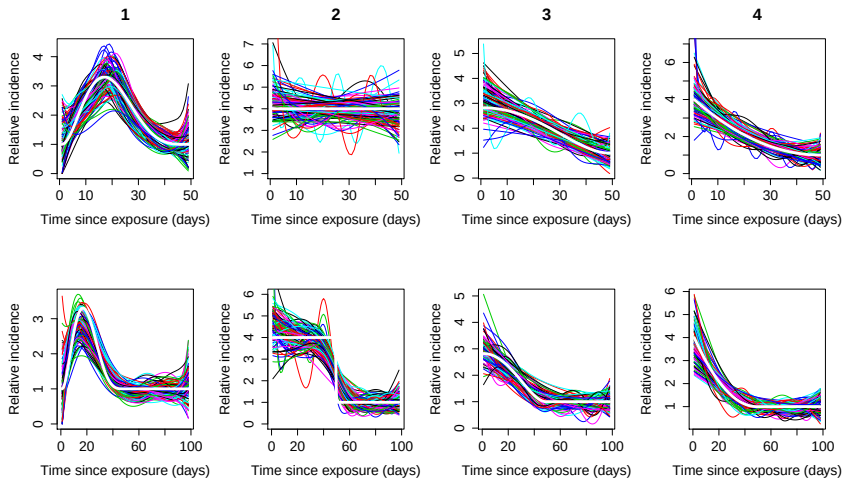


Fig. 5. Estimated relative incidence curves obtained from fitting spline-based SCCS to 100 randomly selected samples with the true relative incidence function in thick white. Top row: 49 day nominal risk period; bottom row: 98 day nominal risk period.

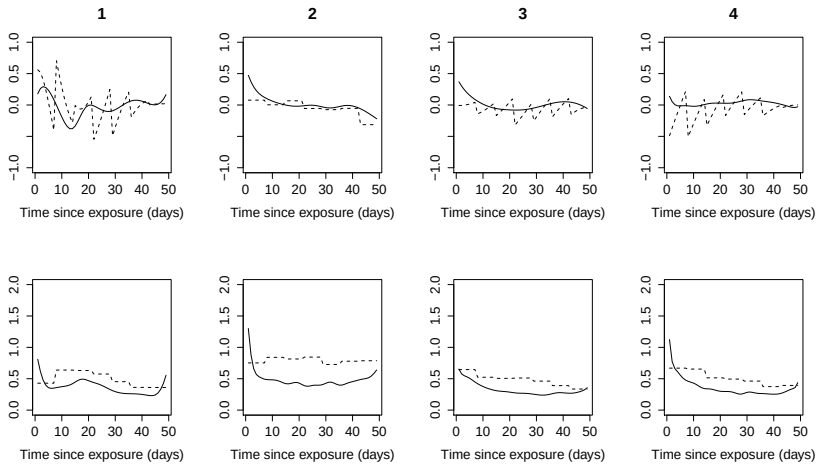


Fig. 6. Bias (top row) and standard deviation (bottom row) of estimates obtained by fitting spline-based SCCS (solid lines) and standard SCCS (dotted lines) to the simulated data sets.

Table 1. Relative incidence (RI) estimates of exposure to MMR vaccine and 95% variability bands obtained from fitting parametric SCCS method with varying exposure groups and 21 age groups.

Risk periods (Days)	Relative Incidence (RI)	95% variability bands Upper	Lower
	Analysis (1)		
(0, 6]	1.23	0.92	1.63
(6, 11]	3.49	2.88	4.23
(11, 18]	0.83	0.60	1.14
(18, 22]	1.49	1.09	2.05
(22, 26]	1.09	0.75	1.58
(26, 30]	1.17	0.82	1.67
(30, 36]	1.16	0.86	1.56
(36, 40]	1.37	0.98	1.91
(40, 45]	0.98	0.69	1.39
(45, 50]	1.10	0.79	1.54
	Analysis (2)		
(0, 11]	2.25	1.91	2.65
(11, 30]	1.09	0.91	1.30
(30, 50]	1.14	0.96	1.35
	Analysis (3)		
(0, 25]	1.62	1.42	1.86
(25, 50]	1.18	1.01	1.39
(50, 75]	1.17	0.99	1.37

Table 2. Mean integrated square error (MISE) and standard deviation (SD) obtained from spline-based and standard SCCS models. Each simulated data set was fitted twice by the two methods with nominal risk periods of 49 and 98 days; age effects are step functions.

	Spline-based SCCS		Standard SCCS with groups of length 7 days		Standard SCCS with groups of length 49 days	
Scenario	MISE	SD	MISE	SD	MISE	SD
			Potential risk perios of 49 days			
1	7.98	5.69	14.99	8.20	37.93	3.49
2	9.58	10.19	31.37	16.43	5.50	7.56
3	5.45	5.63	12.34	6.21	22.39	2.92
4	6.48	8.38	14.65	7.30	43.49	4.59
			Potential risk period of 98 days			
1	14.88	7.10	20.07	7.93	38.12	3.41
2	34.11	13.75	38.75	18.55	8.01	10.13
3	6.44	5.28	20.00	18.79	22.66	2.66
4	8.15	6.82	19.04	8.20	44.23	3.06

Table 3. Mean integrated square error (MISE) and standard deviation (SD) obtained from spline-based and standard SCCS models, with 95% coverage probabilities of the relative incidence value at 10, 25, and 45 days since start of exposure; Age effects are continuous.

			95% Coverage probability		
Scenario	MISE	SD	10	25	45
Spline-based SCCS method					
1	7.76	5.12	90	96	95
2	10.36	8.45	96	98	96
3	5.42	5.29	96	97	95
4	7.63	9.52	96	93	94
5	8.56	8.20	97	97	93
6	12.67	11.07	95	95	94
Standard SCCS method					
1	12.91	6.33	93	96	93
2	23.06	12.41	96	93	97
3	10.60	5.40	93	97	99
4	11.22	5.15	98	95	93
5	13.96	7.34	93	99	96
6*	8.25	6.94	92	92	95

* the standard SCCS model was fitted with three exposure groups